DOCUMENT RESUME

ED 307 293 TM 013 293

AUTHOR Plake, Barbara S.; And Others
TITLE Factors Influencing the Degree of Intrajudge Consistency during the Standard Setting Process.
PUB DATE Mar 89
NOTE 18p.; Paper presented the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, March 28-20, 1989).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cutting Scores; Evaluation Methods; Evaluative Thinking; *Evaluators; *Interrater Reliability; Item Analysis; Minimum Competencies; Rating Scales; *Standards
IDENTIFIERS *Standard Setting

ABSTRACT

The accuracy of standards obtained from judgmental methods is dependent on the quality of the judgments made by experts throughout the standard setting process. One important dimension of the quality of these judgments is the consistency of the judges' perceptions with item performance of minimally competent candidates. Several interrelated sources of intrajudge consistency can be identified. Variables related to the judges themselves may provide a source of differential consistency of ratings during the standard setting process. Aspects of the items within the examination evaluated in the standard setting process may be an additional contributing source of intrajudge consistency. These two types of factors may interact to yield an additional source of inaccuracy in the judges' ratings. Finally, dimensions of the cutoff study could affect the degree of intrajudge consistency. Strategies to improve intrajudge consistency include: (1) periodic rating; (2) utilizing estimates of minimally competent item performance; (3) analyzing empirical data on item performance; (4) reviewing mathematical models for estimates of item performance; and (5) providing descriptive data on judge groups' ratings. Research is needed to demonstrate the potential power of these various techniques. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED307293

Factors Influencing the Degree of Intrajudge Consistency During the Standard Setting Process

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
BARBARA S. PLAKE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Barbara S. Plake

University of Nebraska-Lincoln

Gerald J. Melican

Craig N. Mills

Educational Testing Service

Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1989

TM013293

BEST COPY AVAILABLE

Abstract

The accuracy of standards obtained from judgmental methods is dependent on the quality of the judgments made by experts throughout the standard setting process. One important dimension of the quality of these judgments is the consistency of the judges perceptions with item performance of minimally competent candidates. This paper reviews factors that may influence intrajudge consistency and then makes recommendations about possible strategies to monitor and maintain rater consistency throughout the standard setting process. Research is needed to empirically demonstrate the potential power of these various techniques for improving and maintaining intrajudge consistency.

Factors Influencing the Degree of Intrajudge Consistency During the Standard Setting Process

The cutoff score values resulting from employing judgmental standard setting methods are subject to several sources of error. The quality of these ratings provided by the expert judges is critical to the appropriateness of the resultant standard. van der Linden (1982) discussed three sources of arbitrariness in judgmental standard setting practices including intrajudge inconsistency. Intrajudge inconsistency results from inconsistencies between a judge's rating of the difficulty of the test items and the probability that a minimally competent candidate will correctly answer each item.

Intrajudge inconsistency and possible methods of minimizing this source of error are the topics of this paper. The other two sources of invalidity of judgmental standards setting methods discussed by van der Linden are: (a) differences in the conceptualizations of what constitutes mastery among the various judgmental standard setting techniques, and (b) differences in interpretations of the learning objectives among the expert judges. The first of these two sources is of concern during the planning stages of the study and, although this is an extremely important concern, for the purpose of this paper, it is assumed that the appropriate choice of method has been made. The second of the two sources, different interpretations of the learning

objectives, is one of the objects of the training phase of a standard setting study and, again, it is assumed in this paper that the training has been adequately performed.

The purposes of this paper are to: (a) review factors that may influence the degree of intrajudge consistency throughout a standard setting procedure, and (b) make recommendations about possible strategies to improve intrajudge consistency.

Factors That May influence the Degree of Intrajudge Consistency

Several interrelated sources of intrajudge consistency can be identified. First, variables related to the judges themselves may provide a source of differential consistency of judge's ratings during the standard setting process. Second, specific aspects of the items within the examination evaluated in the standard setting process may be an additional contributing source of intrajudge inconsistency. Third, factors related to the judges themselves may interact with those related to the examination to yield an additional source of inaccuracy in the judges' ratings. Finally, dimensions of the cutoff study procedures could also affect the degree of intrajudge consistency.

Factors Related to Judges

Focusing on the judges themselves, several potential sources of inconsistency may be postulated. The group of judges usually

differ in life experiences, speciality, and entry and continuing professional skills. Often, in fact, the judges are selected specially to represent a distribution of such background characteristics. It is possible that these background factors may affect differentially the item ratings in a specific content portion of the examination. Research investigating the effects of rater expertise has shown some inconsistency across rater groups depending on rater background (Jaeger, 1982; Norcini, Shea, & Kanya, 1988). However, neither of these studies directly considered the possibility of erosion of rater consistency as a potential reason for different standards achieved across different rater groups.

The individual raters may also have idiocynratic perceptions about the skills and competencies of the minimally competent candidate for specific content or processing level. For example, one rater may evaluate items as difficult because the content, while critical to minimum level of competent performance, is infrequently required in practice. While a generic discussion of the skills and competencies of the minimally competent candidate is part of the training of the judges, it is probable that some systematic differences in perception of specific content and processing skills will still be a source of inconsistency during the item rating process. This may be a problem particularly in an examination with a broad range of skills and competencies.

In addition, inaccuracies in the judges' ratings may be accentuated by error caused by fatigue or concentration lapses during the item rating process. Long examinations or rating sessions could exacerbate this source of error.

Factor Related to the Examination

As the examination serves as the stimulus for obtaining ratings by the judges on the predicted performance of the minimally competent candidate, another obvious source of rating inaccuracy is the examination itself. Previous research has considered the possibility of the influence of test specific variables on judgments obtained during a standard setting process. The overall difficulty of the test and the context of the item (location and proximity to easy or difficult items) do not appear to be disruptive to the rater consistency across rating occasions (Melican, & Plake, in press). Other research has concentrated on identifying item types or other examination dimensions that may influence the quality of judgments obtained during a standard setting study (Jaeger, 1982; Melican, Mills, & Plake (in press); Melican, & Thomas, 1984). None of these studies, however, have specifically concentrated the impact of these item dimensions on consistency of rater perceptions during the standard setting process.

Additionally, raters' perceptions of the examination quality or utility for certification/licensure decisions could also be

a contributor to lack of consistency in rating perceptions of item difficulty. For example, if raters disagree with the validity of the examination for certification/licensure decisions, less serious effort or concentration during the item rating procedure could yield inaccurate item ratings.

Factors Related to the Interaction Between Judges and Examination

It is possible that there could be an interaction between rater variables and examination variables in their impact on rater consistency. For example, more novice raters may be differentially susceptible to increases in the degree of rating inconsistency for long examinations or examinations with higher proportions of complex or demanding items.

Factors Related to Cutoff Study Procedure

In addition to influences of the judges, the examination, and possible combinational influences of certain judges with specific examination variables on the degree of consistency of judges ratings, dimensions of the cut score study procedures may also impact the ability of the judges to provide consistent item ratings throughout the standard setting process. The number of examinations to be rated, the comfort and confidence of the raters with the method used to achieve the ratings, the nature and context of the group discussions, the presence or absence of retraining between ratings, are some of the dimensions of cutoff study procedures that could affect the ability of raters to

provide consistent ratings throughout the standard setting process.

Therefore, there are many factors that may contribute to inaccuracies of item ratings during the standard setting study. It is important that these sources of intrajudge rating errors are identified, minimized, and monitored during the entire standard setting process.

Strategies to Improve Intrajudge Consistency

There are several possible ways that rater consistency could be monitored during the standard setting process. These include providing periodic retraining of the judges, utilizing estimations of minimally competent candidates item performance in the rating process, and providing judges with information about the other judges' ratings.

Periodic Rating

One method to attempt to maintain an acceptable degree of intrajudge consistency during the standard setting process is to interrupt the rating task with additional group discussion and possible retraining. While this would help to insure maintainence of the judges perception of the competencies of the minimally competent candidates, there are several potential drawbacks to this approach. First, the overall time to accomplish the standard setting study would be increased. Of course, efficiency is only relevant when quality of information

is maintained, so time efficiency is not the major deterent to this approach. Rater fatigue caused by increased rating time, disruption of rater concentration, and the increased possibility of social influence effects by dominant committee members (Fitzpatrick, 1984; Mills, 1984; Mills, & Peck, 1983), could serve to produce more negative than positive outcomes from the periodic interruption and discussions during the standard setting process.

Utilizing Estimations of Minimally Competent Item Performance

Since rater inconsistency is defined as a function of the discrepancy between the judges subjective estimation of the probability that a hypothetical minimally competent candidate will correctly answer the item and the actual probability of a minimally competent candidate getting the item right, one way to increase intrajudge consistency is incorporate information about examinees on the test items. This information about minimally competent candidates' item performance could be based on either empirical item difficulties or on mathematical models which yield probabilities of correct item performance by candidates whose ability level is at the minimum level for competent performance.

Empiricial Data on Item Performance. One possibility of monitoring the degree of intrajudge consistency would be to involve previous data on the probability that examinees can correctly answer the test items if this data is available. Since the cut score study precedes the operational administration of

the examination, sometimes appropriate item performance data is not available. However, pilot data from item tryouts or preadministered items in an item bank could serve as the basis for empirical item data. One issue in the use of empirical data is on what examinee group should the empirical data be based. Some applications of this approach have utilized data from the total group of examinees (Cross, Impara, Frary, & Jaeger, 1984; Jaeger, 1978; Norcini, Shea, & Kanya, 1988). While it is questionable whether this is the appropriate norm group (Mills, & Melican, 1988), the use of only a minimally competent group raises other issues such as accuracy of estimation due to smaller sample sizes and how the "empirical" minimally competent candidates should be operationally defined.

If empirical data are used then another issue is whether to make the data transparent to the raters. One application of empirical data in the item rating process involves presenting to the raters the previous item performance information as part of the item rating process (Cross, Impara, Frary, & Jaeger, 1984; Jaeger, 1978; Norcini, Shea, & Kanya, 1988) as a means for a "reality check" for the individual raters consideration. Other possibilities for use of empirical item data entail embedding items in the examination with previous item performance data in with the operational test items (Lorge, & Kruglov, 1953; Thorndike, 1982). Rater judgements for the items with historical

item difficulty information is then checked for the degree of rater inconsistency. Sometimes this is used to simply evaluate the overall accuracy of the judges before deciding whether to incorporate all the judges raters in the algorithm for establishing the cutoff score value. However, evidence of rater inconsistency could also be used to flag possible rating inaccuracies during the item rating process. Under this approach, a mechanism is needed to monitor and identify when a "drift" in the degree of rater consistency is occuring. One technique could be to plan for periodic interruption during rating process by the administrator to allow for an assessment of the degree of continued intrajudge consistency. This approach has the same potential for negative impact for disruption as was found during the rating process as identified earlier. A more desirable technique would be to have the raters achieve their item rating on a microcomputer so that the relationships between rater's judgement and empirical data could be monitored on-line and algorithms could be preprogrammed to flag unacceptable deviations in degree of rater consistency.

Mathematical Models for Estimates of Item Performance

This method, suggested by van der Linden in 1982, involves using estimates of the probability that the minimally competent candidate will correctly answer the item from mathematical models of item performance (such as Item Response Theory) instead of

empirical group data to compare with judge's item performance rating. Microcomputer technology could be used to: (a) estimate the probability the minimally competent candidates would correctly answer the item using precalibrated item parameters, (b) monitor the relationship between rater estimates of item difficulty and the estimates based on the mathematical models, and (c) based on preprogrammed decision rules, flag instances where the discrepancy between the judge's subjective estimate and the mathematical model estimate exceed a minimal tolerance level. The system could also be programmed to present on-line training or review to reestablish accceptable levels of intrajudge consistency.

Providing Descriptive Data on Judge Groups' Ratings

Yet another approach to maintain intrajudge consistency during the rating process would be to provide the group of judges with information about the distribution ratings of the entire rater team. This information could be used simply for informational purposes for the judges in the item assessment process or some expectation for consensus may be imposed. Based on a Delphi-type technique, this type of information procedure involving group rating has been utilized by Jaeger (1982), Melican and Mills (1987) and Melican, Mills, and Plake (1988). This technique has the additional potential benefits of providing a more flexible means for obtaining the judges' ratings (as the

rating could be circulated through the mail), and allowing for the opportunity for the judges to alter previous ratings based on the collective information of the group of judges.

Conclusions

Intrajudge inconsistency is a source of inaccuracy in standards derived by judgmental standard setting methods. This paper has reviewed several sources of intrajudge inconsistency and has presented approaches designed to improve the accuracy of ratings obtained throughout the standard setting process. These methods differ in their approach to assessing and/or monitoring the consistency of the judges. Procedures for periodic interruption of the standard setting process in order to provide feedback to the judges and for on-line monitoring of the consistency of judges' ratings were discussed. Of these two approaches, feedback via computerized communication was viewed potentially as less disruptive, less time consuming, and less susceptive to influence by dominant committee members.

The methods for improving rater consistency also differ in the source of item information provided to the raters in communicating information about the probability that examinees will get the item correct for comparison with judges' estimations of this probability. This comparison could either be based on empirical data or estimates based on mathematical models of item performance by minimally competent candidates.

Other strategies for improving degree of rater consistency involve communicating to the judges information about the other judges' ratings of these items. A hybrid of these methods could be to give raters both information about performance of relevant examinee groups and information about the ratings by the other panel members.

Additional research is needed to empirically demonstrate the potential power of these various techniques for improving intrajudge stability. It is important that professionals involved in the design, training, implementation, and application of standard setting studies recognize that it is not sufficient to simply train judges to an acceptable level of consistency prior to the standard setting process. It is also important to monitor and maintain that level of consistency throughout the item rating process. The validity of the standards set by these judgmental standard setting methods depend on consistent ratings obtained throughout the setting study.

References

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards in the National Teachers Examinations. Journal of Educational Measurement, 21, 113-129.

Fitzpatrick, A. R. (April 1984). Social influence in standard setting: The effect of group intraction on individual's judgments. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Jaeger, R. M. (Spring 1978). A proposal for setting a standard on the North Carolina High School Competency Test. Paper presented at the meeting of the North Carolina Association for Research in Education, Chapel Hill, NC.

Jaeger, R. M. (1982). An iterative structured judgement process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4, 461-475.

Lorge, I., & Kruglov, L. K. (1953). The improvement of the estimates of test difficulty. Educational and Psychological Measurement, 13, 34-46.

Melican, G. J., & Mills, C. N. (April 1987). The effect of knowledge of other judges of item difficulty in an interative process using the Angoff and Nedelsky methods. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Melican, G. J., & Thomas, N. (April 1984). Identification of items that are hard to rate using Angoff's standard setting method. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Melican, G. J., Mills, C. N., & Plake, B. S. (April 1988). Accuracy of option elimination prediction: Effect of a technique to improve rater agreement with empirical data. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Melican, G. J., Mills, C. N., & Plake, B. S. (in press). Accuracy in item performance predictions based on the Nedelsky standard settings method. Educational and Psychological Measurement.

Mills, C. N. (April 1984). Social issues in the design of standard setting studies in educational settings. American Educational Research Association, New Orleans.

Mills, C. N., & Melican, G. J. (1988). Estimating and adjusting cutoff scores: Features of selected methods. Applied Measurement in Education, 1, 261-275.

Mills, C. N., & Peck, H. I. (May 1983). Standard setting: Theoretical, practical, and situational concerns. Invited paper presented at the 4th Annual Southwestern Conference on Testing Issues, Dallas.

Norcini, J. J., Shea, J. A., & Kanya, D. T. (1988). The effect of various factors on standard setting. Journal of Educational Measurement, 25, 57-65.

Plake, B. S., & Melican, G. J. (in press). Effects of item content on intrajudge accuracy of expert judgments via judgmental standard setting methods. Educational and Psychological Measurement.

Thorndike, R. L. (1982). Item and score conversion by pooled judgment. In Holland, P. W. and Rubin, D. B. (Eds.). Test Equating, Academic Press, New York.

van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. Journal of Educational Measurement, 19, 295-307.